A Second Wave of Network Technologies

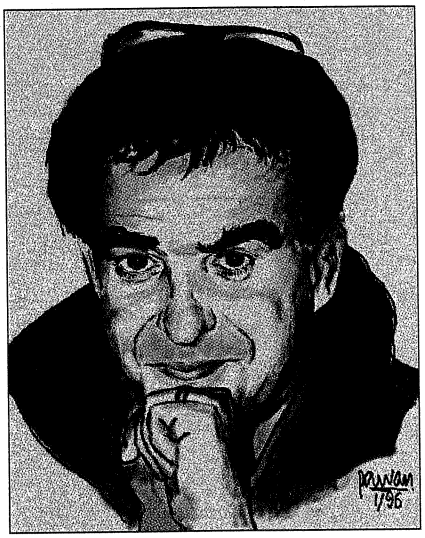
Statistical learning promises to help manage the data glut.

In order for the Internet and intranets to evolve into a highly productive environment for science, engineering, and business, we must develop a new set of technologies. In the first technology wave, the primary problem has been connectivity and bandwidth. Hardware, such as fibers and routers, and software, such as protocols and caching, comprise the Internet's core technologies today.

These technologies provide as many bits as possible, as fast as possible. Even as we solve connectivity problems, however, a new challenge already faces network users: How do you find and manage the vast amount of information on local file systems or the Web? The second Internet wave will include technologies for searching, indexing, and organizing huge amounts of data to support decision making, e-commerce, education, and research. These technologies must "understand" documents, images, and movies, as well as more structured objects such as spreadsheets, parts catalogs, GIS layers, libraries of molecules, genetic sequences, and so on.

> "How do you find and manage the vast amount of information on local file systems or the Web?"

The understanding doesn't need to be complete, just sufficient to do the job. However, the understanding must be able to adapt to the Web's rapidly changing structure and the user's evolving needs. I believe a rapidly evolving new set of techniques, based on a new area of applied mathematics called statistical learning, will be a key technology of this second wave. You can apply these methods



About the Author:
***Tomaso A. Poggio**, Ph.D., is Uncas and Helen Whitaker Professor in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at MIT, and co-director for the Center for Biological and Computational Learning. His present work is motivated by the belief that machine learning is at the core of developing intelligent machines and understanding the brain. E-mail him at tp@ai.mit.edu.*

to develop classifiers that extract data items, such as paragraphs in documents or desired parts of images and videos, by assigning them to meaningful classes. You can also use these methods to extract information—including relationships between data items.

For example, in the future, you might be able to use these technologies to automate document catalogs maintenance, such as those currently developed manually by Yahoo! and other Web search engines. You could also apply them to sort e-mail into relevant categories, identify junk mail, and route documents automatically to the appropriate people—even provide automatic answers. Unlike existing text-based information-retrieval methods, you can apply these methods equally well to text, images, and visual and audio recordings, as well as to numerical and scientific-data objects such as genetic sequence data, galactic spectra, and so on. Finally, you can customize these learning methods by fast training and execute them continually in the rapidly changing Web.

Quote:

"So we went to Atari and said, 'Hey, we've got this amazing thing, even built with some of your parts, and what do you think about funding us? Or we'll give it to you. We just want to do it. Pay our salary, we'll come work for you.' And they said, 'No.' So then we went to Hewlett-Packard, and they said, 'Hey, we don't need you. You haven't got through college yet.' "

— *Apple Computer Inc. founder Steve Jobs on attempts to get Atari and H-P interested in his and Steve Wozniak's personal computer*

> **A skilled human can answer many types of queries existing search engines can't.**

A few computer scientists from the machine-learning community are beginning to develop techniques to manage, organize, and search for multimedia digital information by harnessing the new statistical learning theories and algorithms. For instance, a small consortium of researchers at MIT and in Oregon and Illinois expect to develop prototype systems to classify and route e-mail messages and to search, categorize, and extract information on the Web.

Consider a typical, small-scale problem this research project will tackle. The MIT Admissions Office typically receives 500 e-mail messages per day that are answered by two full-time employees. Imagine a system that could automatically classify e-mail messages according to an evolving set of categories. Many of these messages contain standard questions: The system could automatically answer those messages with canned text, and route others to specific people automatically. As admissions policies, personnel, and course requirements change, the system would adapt its classifications with a modest amount of human interaction.

Commercial systems with primitive capabilities of this type are becoming available, but we need to develop technologies to improve their limited performance. Consider a second, rather obvious example. Commercial search engines for the Web continue to improve, but they can't answer many types of complex, multimedia queries. A skilled human can answer many types of queries existing search engines can't. For example, consider these requests: (a) show pictures of students currently supervised by Professor Tom Dietterich; (b) retrieve images of portable telephones manufactured by Motorola in 1998. You can answer the first query by finding Dietterich's home page, tracing links from that page to the home pages of his students, and analyzing images on those pages to determine whether they're images of the student, as opposed to images of his or her spouse and children. You could answer the second query by searching Web pages or catalogs at Motorola, or online magazine articles for old product reviews.

> **Imagine a system that could automatically classify e-mail messages according to an evolving set of categories.**

Needless to say, these kinds of queries are becoming more common, and we need to create automatic systems to answer them. It's clear that developing smarter classification software for multimedia data will play a key role in enabling a second, more intelligent wave of Internet technologies. We need automatic techniques to route, organize, and search information to help individuals and organizations exploit the ocean of data that the computer networks are creating—instead of drowning in it. ◆

FAQtoid:

The Osborne 1, created in 1980 by Adam Osborne of Osborne Computer Corp., was the first commercially successful portable personal computer. *And* the first computer to be sold with bundled software packages. But competition took its toll: On September 13, 1983, Osborne Computer Corp. declared bankruptcy.